Non Sequential Recursive Pair Substitution: Some Rigorous Results

Dario Benedetto[1], Emanuele Caglioti[1] and Davide Gabrielli[2]

[1] Dipartimento di Matematica, Università di Roma "La Sapienza", P.le A. Moro 2, 00185 Roma, Italy. e-mail: benedetto@mat.uniroma1.it ; caglioti@mat.uniroma1.it

[2] Dipartimento di Matematica, Università dell'Aquila, via Vetoio Loc. Coppito, 67100 L'Aquila, Italy. e-mail: gabriell@univaq.it

Abstract. We present rigorous results on some open questions on NSRPS, non sequential recursive pairs substitution method (see Grassberger in [4]). In particular, starting from the action of NSRPS on finite strings we define a corresponding natural action on measures and we prove that the iterated measure becomes asymptotically Markov. This certify the effectiveness of NSRPS as a tool for data compression and entropy estimation.

AMS classification scheme numbers: 94A17

Keywords: information theory, source and channel coding, entropy, data compression.

1. Introduction

We consider here a suitable non sequential recursive pair substitution method (NSRPS) which has been proposed by Jimenez-Montaño, Ebeling and others [3]. This method has been studied and precisely defined by P. Grassberger as a tool for data compression and entropy estimation [4]. He deduced some important properties of the method and used it to estimate the entropy of the written English.

In particular the results found in [4] and the conjectures made therein are the main motivation for this paper.

Data compression is one of the most interesting research fields in Information Theory both from the applied and from the theoretical viewpoint. In particular data compression algorithms provide a powerful tool for the measure of the entropy and more in general for the estimation of complexity of a sequence. The first algorithms (Shannon-Fano, Huffman, see for example [1], [5]) were based on the suitable coding of single characters, or of strings of a fixed and small number of characters. A great improvement in the field of data compression has been given by the dictionary-based compression methods LZ77 [7], LZ78 [8] and LZW [6] in which variable-length strings are suitably encoded. In particular in LZ78 a sequence is encoded as a list of phrases. Initially the phrases coincide with the characters and then any new phrase is obtained sequentially by adding a character to one of the existing phrases. The NSRPS method we

are going to study here, even if different in many respects from this dictionary methods, has some similarity with LZ78 and in particular with a variation of LZ78 which has been recently proposed [2].

The NSRPS method works in the following way. Let us consider a sequence $\underline{s}^0$ built with the characters of a finite alphabet $A = \{a_0, ..., a_{m-1}\}$. For any given i, j let n_{ij} be the number of non-overlapping occurrences of the string $a_i a_j$ in $\underline{s}^0$, and let be i_0, j_0 the pair (or one of the pairs) for which n_{ij} is maximum. Now let us define a new sequence $\underline{s}^1$ obtained from $\underline{s}^0$ by substituting any occurrence of the pair $a_{i_0} a_{j_0}$ with a new symbol a_m. The new sequence is shorter than the previous one and its alphabet has one character more. Then starting from $\underline{s}^1$ we define a new sequence $\underline{s}^2$ with the same procedure, et cetera. We call a single step of NSRPS a "pair substitution" (the one for example that transforms $\underline{s}^0$ into $\underline{s}^1$).

For sake of clearness let us consider two specific examples when the initial sequence is binary. For first let us consider the case in which

$$\underline{s}^0 = 0010101010001001010101110101.....$$

and we substitute 01 with the new character 2. We obtain

$$\underline{s}^1 = 02222002022221122.....$$

As said above the sequence $\underline{s}^1$ is shorter then $\underline{s}^0$. In particular, denoting with $|\underline{s}|$ the length of a generic sequence $\underline{s}$, we have

$$|\underline{s}^1| = |\underline{s}^0| - \#\{01 \subseteq \underline{s}^0\},$$

where $\#\{01 \subseteq \underline{s}^0\}$ is the number of times we find 01 in the string $\underline{s}^0$. Dividing by $|\underline{s}^0|$

$$\frac{|\underline{s}^1|}{|\underline{s}^0|} = 1 - \frac{\#\{01 \subseteq \underline{s}^0\}}{|\underline{s}^0|}.$$

We always work with sequences extracted by an ergodic measure μ. Then taking the limit as $|\underline{s}^0| \to \infty$ we get, for almost all sequences $\underline{s}^0$, that

$$\frac{1}{Z} := \lim_{|\underline{s}^0| \to \infty} \frac{|\underline{s}^1|}{|\underline{s}^0|} = 1 - \mu(01). \tag{1.1}$$

Another important fact to notice is that the transformation is invertible (see Section 2), and then the amount of information of the two sequences is the same (see Section 3). Therefore, if $h(\underline{s})$ is the entropy per character of $\underline{s}$:

$$h(\underline{s}^0) = \frac{h(\underline{s}^1)}{Z}.$$

The second example we consider is when the pair to be substituted is made by two equal characters. Let us consider the sequence

$$10011001000000110010000010000010001\ldots$$

and let us substitute 00 with 2. We found the new sequence

$$121121222112122012212 01\ldots$$

The main difference with the case considered before is the fact that in this case we do not substitute with 2 all the pairs of consecutive 0 in $\underline{s}^0$. For instance $1001 \to 121$, but $10001 \to 1201$. It is easy to deduce that in this case (1.1) changes in

$$\frac{1}{Z} = 1 - \mu(00) + \mu(000) - \mu(0000) + \mu(00000) - \ldots\ldots \tag{1.2}$$

This example shows that under a NSRPS the probabilities of strings can behave in a complicate way. In spite of this fact, the substitution process transform a Markov sequence in a Markov sequence, as proved by Grassberger in [4].

In general, if the starting sequence is not Markov it does not becomes Markov after a finite number of transformations. Nevertheless it is reasonable to expect that the sequences tends to become Markov as the number of transformations tends to infinity. This is exactly what was conjectured in [4] and what we prove here.

More precisely the main facts we prove are the following.

In any pair substitution the conditional entropy h_1 (i.e. the entropy of a character conditioned to the previous character), suitably normalized, does not increase. If the process is already Markov then it stays constant (truly, there are other rare cases in which h_1 stays constant, see Section 5 and Section 8).

This is a general property of the pair transformations and holds true whatever is the substitution made. An immediate corollary of this fact is that Markov sequences are transformed in Markov sequences.

As the number of transformations goes to ∞ and also the inverse of the average shortening Z diverges, the (suitably normalized) conditional entropy h_1 tends to the entropy of the sequence. In this sense we prove that in the limit the process becomes Markov. In particular this is the case if any time we substitute the pair of characters which maximizes the number of nonoverlapping occurrences. This condition is not strictly necessary but, as we shall see in Section 5, not for all the sequences of substitutions it holds the result.

The paper is organized as follows.

In section 2 we will fix notations and give some preliminary results. In particular we will discuss how pair substitutions act on strings and give a natural definition of a corresponding action on ergodic measures.

In section 3 we will state results on how pair substitutions act on entropies.

In section 4 we prove the main result of the paper.

In section 5 we discuss some examples.

In section 6 we give some concluding remarks.

In sections 7, 8 we collect technical results on measures and entropies transformations under the action of a pair substitution, respectively.

2. How pair substitutions act on strings and measures

2.1. Strings

Given an alphabet A we denote with $A^* := \cup_{k=1}^{+\infty} A^k$ the set of finite words in the alphabet A. Elements of A^* are indicated with underlined lower case Latin letters $\underline{w}, \underline{x}$, *etc.* The same notation will be used also for infinite (elements of $A^{\mathbb{N}}$) and double infinite words (elements of $A^{\mathbb{Z}}$). An element $\underline{w}$ has length $|\underline{w}|$ and, if $|\underline{w}| = k$, it is also indicated with $w_1^k := w_1 \dots w_k := (w_1, \dots, w_k)$.

Let us consider $x, y \in A$ (including $x = y$), $\alpha \notin A$, and $A' = A \cup \{\alpha\}$. A *pair substitution* is a map $G = G_{xy}^{\alpha} : A^* \to A'^*$ which substitutes ordinately the occurrence of xy with α. More precisely $G\underline{w}$ is defined by substituting in $\underline{w}$ the first occurrence from the left of xy with α, and then repeating this procedure till the end of string.

We define also the map $S = S_{\alpha}^{xy} : A'^* \to A^*$, which acts on the words $\underline{z} \in A'^*$ substituting any occurrence of the symbol α with the pair xy.

Notice that the map G is injective and not surjective while the map S is surjective and not injective. Notice also that $S|_{G(A^*)} = G^{-1}$, i.e.

$$S(G(\underline{w})) = \underline{w} \quad \text{for any} \quad \underline{w} \in A^*. \tag{2.3}$$

We remark that these definitions work also in the case of infinite sequences $\underline{w} \in A^{\mathbb{N}}$ and $\underline{z} \in A'^{\mathbb{N}}$.

It is easy to see that the set of admissible words $G(A^*)$ is a subset of A'^* which can be described by constraints on consecutive symbols: in the case $xy \to \alpha$, with $x \neq y$, $G(A^*)$ consists of the strings of A'^* in which does not appear the pair xy; in the case $xx \to \alpha$, $G(A^*)$ consists of the strings of A'^* in which do not appear the pairs xx and $x\alpha$. An important fact is that after the application of more pair substitutions, the set of admissible words remains described by constraints on consecutive symbols. This follows from the fact that a pair substitution maps pair constraints in pair constraints, as stated in the following theorem.

Theorem 2.1 *Let $\{V_{a,b}\}_{a,b \in A}$ be a matrix with 0–1 valued elements (the constraint matrix), and let A_V^* be the subset of A^* whose elements $\underline{w}$ verify*

$$\prod_{i=1}^{|\underline{w}|-1} V_{w_i, w_{i+1}} = 1,$$

(A_V^* is the set of admissible strings with respect to the pair constraints given by V). There exists a constraint matrix V' with index in A' such that

$$G(A_V^*) = A'^*_{V'}.$$

The proof follows from direct inspection. Here we only write V' in terms of V. Let $z, w \in A \setminus \{x, y\}$: the values of the elements of V' are given by the following tables:

if $x \neq y$	x	y	w	α
x	$V_{x,x}$	0	$V_{x,w}$	$V_{x,x}$
y	$V_{y,x}$	$V_{y,y}$	$V_{y,w}$	$V_{y,x}$
z	$V_{z,x}$	$V_{z,y}$	$V_{z,w}$	$V_{z,x}$
α	$V_{y,x}$	$V_{y,y}$	$V_{y,w}$	$V_{y,x}$

if $x = y$	x	w	α
x	0	$V_{x,w}$	0
z	$V_{z,x}$	$V_{z,w}$	$V_{z,x}$
α	1	$V_{x,w}$	1

Note that these expressions hold if $V_{x,y} = 1$ and $V_{x,x} = 1$ respectively; otherwise, and this is a non interesting case, $G(A_V^*) = A_V^*$.

2.2. Measures

We indicate with $\mathcal{E}(A)$ the set of ergodic stationary measures on $A^{\mathbb{Z}}$, the only measures we are interested in. If $\mu \in \mathcal{E}(A)$ we use the shorthand notation $\mu(\underline{w})$ to indicate the value of the $|\underline{w}|$-marginals of μ on the sequence $\underline{w}$.

The maps G_{xy}^{α} and S_{α}^{xy} induce the maps $\mathcal{G} = \mathcal{G}_{xy}^{\alpha} : \mathcal{E}(A) \to \mathcal{E}(A')$ and $\mathcal{S} = \mathcal{S}_{\alpha}^{xy} : \mathcal{E}(A') \to \mathcal{E}(A)$ in the following natural sense. Let $\mu \in \mathcal{E}(A)$ and $\underline{w} \in A^{\mathbb{N}}$ be a frequency typical sequence with respect to μ, and let $\nu \in \mathcal{E}(A')$ and $\underline{z} \in A'^{\mathbb{N}}$ be a frequency typical sequence with respect to ν. The sequence $G\underline{w}$ is typical for an ergodic measure that we call $\mathcal{G}\mu$ and the sequence $S\underline{z}$ is typical for an ergodic measure that we call $\mathcal{S}\nu$.

More precisely, denoting the number of occurrences of a subword $\underline{s}$ in $\underline{r}$ with $\sharp\{\underline{s} \subseteq \underline{r}\} := \sum_{i=1}^{|\underline{r}|-|\underline{s}|+1} \mathbb{I}(r_i^{i+|\underline{s}|-1} = \underline{s})$, where $\mathbb{I}$ is the characteristic function, it holds

Theorem 2.2 *Let $\underline{s} \in A'^*$ then*

$$\mathcal{G}\mu(\underline{s}) := \lim_{n\to+\infty} \frac{\sharp\{\underline{s} \subseteq G(w_1^n)\}}{|G(w_1^n)|} \tag{2.4}$$

exists and is constant μ almost everywhere in $\underline{w}$, moreover $\{\mathcal{G}\mu(\underline{s})\}_{\underline{s}\in A'^*}$ are the marginals of an ergodic measure on $A'^{\mathbb{Z}}$.

In analogous way, let $\underline{r} \in A^*$; then

$$\mathcal{S}\nu(\underline{r}) := \lim_{n\to+\infty} \frac{\sharp\{\underline{r} \subseteq S(z_1^n)\}}{|S(z_1^n)|} \tag{2.5}$$

exists and is constant ν almost everywhere in $\underline{z}$, moreover $\{\mathcal{S}\nu(\underline{r})\}_{\underline{r}\in A^*}$ are the marginals of an ergodic measure on $A^{\mathbb{Z}}$. It holds

$$\mathcal{S}_{\alpha}^{xy}\mathcal{G}_{xy}^{\alpha}\mu = \mu. \tag{2.6}$$

In Section 7 we give the proof of the theorem and of the following propositions (which we use for the main theorem in Section 4); moreover from (2.4) and (2.5) we write the explicit expressions of $\mathcal{G}\mu$ and $\mathcal{S}\nu$ in terms of μ and ν respectively.

Proposition 2.1 *Let Z_{xy}^{μ} be the inverse of the mean shortening, with respect to μ, of a string under the action of G_{xy}^{α} and let $W = W_{\alpha}^{\nu}$ be the mean lengthening, with respect to ν, of a string under the action of S_{α}^{xy}.*

$$\textit{If } x \neq y \quad Z_{xy}^{\mu} := \lim_{n\to+\infty} \frac{n}{|G(w_1^n)|} = \frac{1}{1-\mu(xy)} \qquad (\mu \textit{ a.e. in } \underline{w}). \tag{2.7}$$

$$Z^{\mu}_{xx} := \lim_{n\to+\infty} \frac{n}{|G(w_1^n)|} = \frac{1}{1-\sum_{k=2}^{+\infty}(-1)^k \mu(\underline{x}^k)} \qquad (\mu \text{ a.e. in } \underline{w}), \tag{2.8}$$

where $\underline{x}^k$ is the sequence of k times x.

$$W^{\nu}_{\alpha} := \lim_{n\to+\infty} \frac{|S(z_1^n)|}{n} = 1 + \nu(\alpha) \qquad (\nu \text{ a.e. in } \underline{z}). \tag{2.9}$$

Moreover

$$W_{\alpha}^{\mathcal{G}^{\alpha}_{xy}\mu} = Z^{\mu}_{xy} \tag{2.10}$$

Proposition 2.2 *Let $\underline{r} \in A^*$, the value of $\mathcal{S}\nu(\underline{r})$ depends only on the values of $\nu(\underline{s})$ with $|\underline{s}| \leq |\underline{r}|$*

We remark that this assertion if false for $\mathcal{G}\mu$, and, in the case of $x = y$, $\mathcal{G}\mu(\underline{s})$ can involve the probability of infinitely many strings of increasing lengths (see Eq.s (7.33)).

Proposition 2.3 *(Invertibility of $\mathcal{S}\nu$)*
If $\nu \in \mathcal{E}(A')$ respects the pair constraints given by G, i.e. for $\underline{z} \in A'^*$

$$\nu(\underline{z}) = 0 \quad \text{if} \;\; \underline{z} \notin G(A^*),$$

then

$$\nu = \mathcal{G}\mathcal{S}\nu.$$

3. How pair substitutions act on the entropy per symbol

Given $\mu \in \mathcal{E}(A)$, $n \geq 1$, and indicating with log the base 2 logarithm function,

$H_n(\mu) := -\sum_{\lvert\underline{z}\rvert=n} \mu(\underline{z}) \log \mu(\underline{z})$	is the n–block entropy,
$h_n(\mu) := H_{n+1}(\mu) - H_n(\mu)$	is the n–conditional entropy,
$h(\mu) := \lim_{n\to+\infty} \frac{H_n(\mu)}{n} = \lim_{n\to+\infty} h_n(\mu)$	is the entropy of μ.

It holds:

$$h(\mu) \leq \ldots \leq h_j(\mu) \leq h_{j-1}(\mu) \leq \ldots \leq h_1(\mu) \leq H_1(\mu). \tag{3.11}$$

Denoting with $\mu(\underline{z}|\underline{w}) := \mu(\underline{w}\,\underline{z})/\mu(\underline{w})$ the conditional probabilities, we say that μ is a $k-$Markov measure if for any $n > k$, $\underline{w} \in A^n$ and $a \in A$, $\mu(a|w_1^n) = \mu(a|w_{n-k+1}^n)$. In this case $h(\mu) = h_j(\mu) \quad \forall j \geq k$. We remark that $h(\mu) = h_k(\mu)$ implies that μ is a $k-$Markov measure.

We collect here some results on how entropies transform under the action of $\mathcal{G}$. Proofs are postponed to the technical Section 8.

We will use the shorthand $Z = Z^{\mu}_{xy}$, and sometimes $Z^{\mu} = Z^{\mu}_{xy}$ when we need to stress the reference measure.

Theorem 3.1

$$h(\mathcal{G}\mu) = Zh(\mu). \tag{3.12}$$

In fact the information amount of the string $\underline{w}$ is the same of the string $G(\underline{w})$.

Theorem 3.2

$$h_1(\mathcal{G}\mu) \leq Z h_1(\mu) \tag{3.13}$$

Moreover, if μ is a $1-$Markov measure $\mathcal{G}\mu$ is a $1-$Markov measure.

Let us notice here that the second assertion is a consequence of the first: if μ is a $1-$Markov measure

$$h(\mathcal{G}\mu) \leq h_1(\mathcal{G}\mu) \leq Z h_1(\mu) = Z h(\mu) = h(\mathcal{G}\mu). \tag{3.14}$$

Then $h_1(\mathcal{G}\mu) = h(\mathcal{G}\mu)$; this implies that $\mathcal{G}\mu$ is a $1-$Markov measure.

This theorem can also be generalized.

Theorem 3.3

$$h_k(\mathcal{G}\mu) \leq Z h_k(\mu), \tag{3.15}$$

and $\mathcal{G}$ maps $k-$Markov measures in $k-$Markov measures.

4. The main result

Theorem 3.2 asserts, roughly speaking, that the amount of information of $G(\underline{w})$, which is equal to that of $\underline{w}$, is more concentrated on the pairs of symbols, with respect to the case of the original string $\underline{w}$. This fact suggests that a sequence of pair substitutions can transfer all the information in the distributions of the pairs of symbols. To formalize this assertion, let us define recursively:

the alphabets $A_N = A_{N-1} \cup \{\alpha_N\}$ where $\alpha_N \notin A_{N-1}$, with $A_0 = A$;
the maps $G_N = G^{\alpha_N}_{x_N y_N} : A^*_{N-1} \to A^*_N$, where x_N, $y_N \in A_{N-1}$;
the corresponding maps $\mathcal{G}_N = \mathcal{G}^{\alpha_N}_{x_N y_N}$, $S_N = S^{x_N y_N}_{\alpha_N}$, $\mathcal{S}_N = \mathcal{S}^{x_N y_N}_{\alpha_N}$;
the measures $\mu_N = \mathcal{G}_N \mu_{N-1}$, with $\mu_0 = \mu$;
the normalization $Z_N = Z^{\mu_{N-1}}_{x_N y_N}$;
the composed maps

$$\begin{aligned} \overline{G}_N &= G_N \circ \cdots \circ G_1, & \overline{\mathcal{G}}_N &= \mathcal{G}_N \circ \cdots \circ \mathcal{G}_1, \\ \overline{S}_N &= S_1 \circ \cdots \circ S_N, & \overline{\mathcal{S}}_N &= \mathcal{S}_1 \circ \cdots \circ \mathcal{S}_N; \end{aligned}$$

the corresponding normalization $\overline{Z}_N = Z_N Z_{N-1} \cdots Z_1$ (when we need to specify the initial measure we will use the symbol $\overline{Z}^\mu_N$).

In [4] the author chose at any step the pair of symbols with the maximum of the frequency of non-overlapping occurrences. This fact assures the divergence of $\overline{Z}_N$ as we will prove using Theorem 3.2.

Theorem 4.1 *If at any step N the pair $x_N y_N$ is the pair of maximum of frequency of non-overlapping occurrences between the pairs of symbols of A_{N-1} then*

$$\lim_{N \to +\infty} \overline{Z}_N = +\infty \tag{4.16}$$

In this case the hypothesis of the following (main) theorem is satisfied.

Theorem 4.2 *If*

$$\lim_{N\to+\infty} \overline{Z}_N = +\infty \tag{4.17}$$

then

$$h(\mu) = \lim_{N\to+\infty} \frac{h_1(\mu_N)}{\overline{Z}_N} \tag{4.18}$$

Proof of Th. 4.1

Let p_N the maximum of probability μ_{N-1} on the pair of symbols of A_{N-1}. From the definition of Z_N it follows that

$$\overline{Z}_N \geq \overline{Z}_{N-1}\left(1 + \frac{p_N}{2}\right),$$

(the factor 2 appears for the case of substitution of two equal symbols). We can estimate p_N with

$$p_N \geq 2^{-H_2(\mu_{N-1})},$$

where $H_2(\mu_{N-1}) = -\sum_{a,b\in A_{N-1}} \mu_{N-1}(ab)\log\mu_{N-1}(ab)$ is the 2-block entropy. Using Th. 3.2 and that $H_1(\mu_{N-1}) \leq \log(N-1+|A|)$, with $|A|$ the cardinality of A:

$$H_2(\mu_{N-1}) = h_1(\mu_{N-1}) + H_1(\mu_{N-1}) \leq \overline{Z}_{N-1} h_1(\mu) + \log(N-1+|A|).$$

Then

$$\frac{\overline{Z}_N}{\overline{Z}_{N-1}} \geq 1 + \frac{2^{-\overline{Z}_{N-1} h_1(\mu)}}{2(N-1+|A|)}.$$

The sequence $\overline{Z}_N$ is increasing; by absurd, if $\overline{Z}_N$ tends to a constant, from the previous equation $\overline{Z}_N/\overline{Z}_{N-1} \geq 1 + c/(N-1)$, but this implies $\overline{Z}_N \to +\infty$.

Remark: this proof is also valid in the more general case we choose $x_N y_N$ in such a way that

$$\mu_{N-1}(x_N y_N) \geq c p_N,$$

where c is a constant independent on N.

Proof of Th. 4.2

For the composition $\overline{S}_N$ it holds

$$\overline{S}_N(s_1^n) = \overline{S}_N(s_1)\ldots\overline{S}_N(s_n),$$

where $\overline{S}_N(s_i)$ are words in the original alphabet A. Consider $\underline{r} \in A^*$, $|\underline{r}| = k$ and $\underline{s}$ a typical string for μ_N.

$$\mu(\underline{r}) = \lim_{n\to\infty} \frac{\sharp\left\{\underline{r} \subseteq \overline{S}_N(s_1^n)\right\}}{|\overline{S}_N(s_1^n)|} = \lim_{n\to\infty} \frac{\sharp\left\{\underline{r} \subseteq \overline{S}_N(s_1)\ldots\overline{S}_N(s_n)\right\}}{|\overline{S}_N(s_1^n)|}$$

Notice that

$$\begin{aligned}&\sharp\left\{\underline{r}\subseteq\overline{S}_N(s_1)\ldots\overline{S}_N(s_n)\right\}=\textstyle\sum_{g\in A_N}\sharp\left\{\underline{r}\subseteq\overline{S}_N(g)\right\}\sharp\left\{g\subseteq s_1^n\right\}\\&+\textstyle\sum_{p=2}^k\sum_{g_1,\ldots,g_p\in A_N}\sharp\left\{\underline{r}\frown\overline{S}_N(g_1)\ldots\overline{S}_N(g_p)\right\}\sharp\left\{g_1\ldots g_p\subseteq s_1^n\right\}\end{aligned}\tag{4.19}$$

where $\sharp\left\{\underline{r}\frown\overline{S}_N(g_1)\ldots\overline{S}_N(g_p)\right\}$ is the number of occurrences of $\underline{r}$ in the string $\overline{S}_N(g_1)\ldots\overline{S}_N(g_p)$ which start in $\overline{S}_N(g_1)$ and end in $\overline{S}_N(g_p)$. We obtain

$$\begin{aligned}\mu(\underline{r})&=\lim_{n\to\infty}\tfrac{n}{|\overline{S}_N(s_1^n)|}\Bigg(\textstyle\sum_{g\in A_N}\sharp\left\{\underline{r}\subseteq\overline{S}_N(g)\right\}\tfrac{\sharp\left\{g\subseteq s_1^n\right\}}{n}\\&+\textstyle\sum_{p=2}^k\sum_{g_1,\ldots,g_p\in A_N}\sharp\left\{\underline{r}\frown\overline{S}_N(g_1)\ldots\overline{S}_N(g_p)\right\}\tfrac{\sharp\left\{g_1\ldots g_p\subseteq s_1^n\right\}}{n}\Bigg)\\&=\tfrac{1}{\overline{Z}_N}\Big(\textstyle\sum_{g\in A_N}\sharp\left\{\underline{r}\subseteq\overline{S}_N(g)\right\}\mu_N(g)\\&+\textstyle\sum_{p=2}^k\sum_{g_1,\ldots,g_p\in A_N}\sharp\left\{\underline{r}\frown\overline{S}_N(g_1)\ldots\overline{S}_N(g_p)\right\}\mu_N(g_1\ldots g_p)\Big)\end{aligned}\tag{4.20}$$

Let $\mathcal{P}$ be the projection operator that maps a measure μ to its 1-Markov approximation $\mathcal{P}\mu$ and define $\pi_N^j=\mathcal{S}_{j+1}\ldots\mathcal{S}_N\mathcal{P}\mu_N$. In particular we have $\pi_N^0=\overline{\mathcal{S}}_N\mathcal{P}\mu_N$ and $\pi_N^N=\mathcal{P}\mu_N$. It holds

$$\pi_N^N=\overline{\mathcal{G}}_N\pi_N^0.\tag{4.21}$$

In fact the measures π_N^N and μ_N coincide on the pairs of symbols, then $\pi_N^N(\underline{w})=0$ if $\underline{w}\notin\overline{G}_N(A^*)$, as follows from Th. 2.1. Being $\overline{G}_N(A^*)\subseteq G_N(A_{N-1}^*)$ we can apply Proposition 2.3, obtaining

$$\pi_N^N=\mathcal{G}_N\pi_N^{N-1}.\tag{4.22}$$

Now, also π_N^{N-1} and μ_{N-1} coincide on the pairs of symbols (see Proposition 2.2), then we can iterate the procedure till to obtain Eq. (4.21). Note that

$$\overline{Z}_N^{\pi_N^0}=\prod_{j=1}^N(1+\pi_N^j(\alpha_j))=\prod_{j=1}^N(1+\mu_j(\alpha_j))=\overline{Z}_N^\mu,\tag{4.23}$$

in fact π_N^j and μ_j coincide on the pairs of symbols on A_j. Therefore for any k and any $\underline{r}$ of length k:

$$\begin{aligned}&|\pi_N^0(\underline{r})-\mu(\underline{r})|\le\\&\le\tfrac{1}{\overline{Z}_N}\textstyle\sum_{p=3}^k\sum_{g_1,\ldots g_p\in A_N}\Big(\mu_N+\pi_N\Big)(g_1\ldots g_p)\sharp\left\{\underline{r}\frown\overline{S}_N(g_1)\ldots\overline{S}_N(g_p)\right\}\\&\le 2\tfrac{k^2}{\overline{Z}_N}\end{aligned}\tag{4.24}$$

which tends to 0 when $N\to+\infty$. This implies that

$$\lim_{N\to+\infty}h_k(\pi_N^0)=h_k(\mu).$$

In conclusion, for any k

$$h(\mu)=\frac{h(\mu_N)}{\overline{Z}_N}\le\frac{h_1(\mu_N)}{\overline{Z}_N}=\frac{h(\pi_N^N)}{\overline{Z}_N}=h(\pi_N^0)\le h_k(\pi_N^0).$$

We stress that the third step of the previous chain follows from the definition $\pi^{N}_{N} = \mathcal{P}\mu_{N}$ and that the fourth step follows from (4.21) and (4.23).
Taking the limits $N \to +\infty$ and $k \to +\infty$:

$$h(\mu) = \lim_{N\to+\infty} \frac{h_1(\mu_N)}{\overline{Z}_N}.$$

5. Some examples

We consider here a given sequence of pair substitutions which is not obtained with the procedure of the minimization of the length of the new strings, as prescribed in the NSRPS method.

The initial alphabet is $A = \{0, 1\}$. The first pair substitution is $10 \to 2$, the second $20 \to 3$; in general the N-th substitution is $N0 \to N + 1$. Notice that the infinite composition of these substitutions corresponds to the coding procedure that substitute maximal blocks of k consecutive zeros, and the one that precedes them, with the new symbol $k + 1$.

If the initial measure gives positive probability to the pair 11, then the normalization can not diverge, namely for an initial (typical) string of length L, after the transformations there remain at most $\mu(11)L$ symbols.

Let us notice that only the first substitution involves the symbol one, then it is easy to do the following computations:

$$\begin{aligned}
\mu_N(1|1) \;\; &= \mu_1(1|1) = \frac{\mu(11) - \mu(110)}{\mu(1) - \mu(10)} = \mu(1|11),\\
\mu_N(1|11) &= \mu_1(1|11) = \frac{\mu(111) - \mu(1110)}{\mu(11) - \mu(110)} = \mu(1|111).
\end{aligned}$$

If for the initial measure $\mu(1|111) \neq \mu(1|11)$, then $\mu_N(1|11) \neq \mu_N(1|1)$ for any N and $h_1(\mu_N)/\overline{Z}_N$ can not converge to $h(\mu)$ (the limiting process can not be a 1-Markov process).

On the other hand we can consider as initial measure a finite mean renewal process, that is a stationary process for which the distances between consecutive ones are i.i.d. random variables with distribution $\{p_k\}_{k\geq 1}$ and $E^0 = \sum_{j=1}^{\infty} jp_j < \infty$. The entropy of such a process is

$$h(\mu) = \frac{-\sum_{k=1}^{\infty} p_k \log p_k}{E^0}$$

An explicit computation of the marginals of μ_N is not difficult. It follows that

$$Z_N = Z_{N0}^{\mu_{N-1}} = \frac{E^{N-1}}{E^N}, \qquad \overline{Z}_N = \frac{E^0}{E^1}\frac{E^1}{E^2}\cdots\frac{E^{N-1}}{E^N} = \frac{E^0}{E^N},$$

where $E^N = \sum_{j=1}^{\infty} jp_j^N$ and

$$p_j^N = \begin{cases} p_1 + \ldots + p_{N+1} & j = 1 \\ p_{N+j} & j > 1 \end{cases}$$

Note that if we consider the measures μ_N as measures in the alphabet $\mathbb{N}$, then μ_N weakly converges to the product measure with marginals $\{p_k\}_{k\geq 1}$. From this (or by direct computation) we can derive

$$\lim_{N\to\infty}\frac{h_1(\mu_N)}{\overline{Z}_N}=\lim_{N\to\infty}\frac{H_1(\mu_N)}{\overline{Z}_N}=h(\mu)$$

Let us stress that in this case the process becomes independent, then also $H_1(\mu_N)/\overline{Z}_N$ converges to the entropy. This fact is a consequence of the very particular choice of the initial measure. If the distances between consecutive ones are not distributed independently, but, for instance, with a two step Markov process, then $h_1(\mu_N)/\overline{Z}_N$ and $H_1(\mu_N)/\overline{Z}_N$ do not converge to the entropy.

6. Concluding remarks

The main result proved here says that under the action of the NSRPS procedure any ergodic process becomes asymptotically Markov, i.e. $h_1(\mu_N)/\overline{Z}_N \to h$. A natural question is when the process becomes even independent, i.e. $H_1(\mu_N)/\overline{Z}_N \to h$, as for the very specific example discussed in section 5. In our opinion this is a non trivial question, presumably depending on the behavior of the number of forbidden sequences in the iterated measures.

The results of this paper imply also the fact that a NSRPS algorithm can be used to estimate the entropy of an ergodic source starting from a sequence of sufficiently large length, say L. This is done iterating $N(L)$ pair substitutions with $N(L)$ diverging with L sufficiently slow, and then computing the conditional entropy h_1 of the empirical measure of the resulting sequence. An interesting question is how fast $N(L)$ can diverge with L.

Analogously it is possible to define an asymptotically optimal compression algorithm based on NSRPS: iterating a suitable number of times the pair substitution procedure we ends up with an approximatively Markov sequence; this sequence can be compressed by an algorithm which take into account only the pair correlations (like for instance a suitable arithmetic coder). As before, if the number of substitutions diverges with L sufficiently slow then the compression rate converges to h.

In practice, given a sequence of length L, it is not so obvious to decide in an efficient way what is the optimal number of substitution to make. This point is discussed a little bit in [4] and we do not enter in it.

7. Technical results on measures transformations

7.1. Proof of Theorem 2.2

We do not give a formal proof of the theorem but just a sketch of it (more details are in the analogous proof for Proposition 2.1, in the next subsection). The fact that the limits are almost surely constants can be deduced from the strong law of large numbers.

This fact implies the ergodicity of $\mathcal{G}\mu$ and $\mathcal{S}\nu$ (see theorem I.4.2 pag. 44 of [5]). The compatibility conditions for the families of marginals are easily checked. Formula (2.6) is consequence of (2.3).

7.2. Proof of Proposition 2.1

In the case $x \neq y$, we have

$$|G(w_1^n)| = n - \sharp\{xy \subseteq w_1^n\}$$

so that

$$\frac{n}{|G(w_1^n)|} = \frac{1}{1 - \frac{\sharp\{xy\subseteq w_1^n\}}{n}}$$

and the result (2.7) follows from the strong law of large numbers.
In the case $x = y$ we have that

$$|G(w_1^n)| = n - \sum_{k=2}^{n} \sharp\left\{*x^k* \subseteq w_1^n\right\}\left[\frac{k}{2}\right]$$

where [ ] is the integer part and $\sharp\left\{*x^k* \subseteq w_1^n\right\}$ is the number of blocks of exact length k of consecutive x contained in w_1^n ($*$ represent a possible occurrence of a generic letter different from x). It holds

$$\sharp\left\{*x^k* \subseteq w_1^n\right\} = \sharp\left\{x^k \subseteq w_1^n\right\} - 2\sharp\left\{x^{k+1} \subseteq w_1^n\right\} + \sharp\left\{x^{k+2} \subseteq w_1^n\right\}$$

Now we have

$$\frac{n}{|G(w_1^n)|} = \frac{1}{1 - \sum_{k=2}^{n}(-1)^k\left(\frac{\sharp\left\{x^k\subseteq w_1^n\right\}}{n}\right)}$$

that converges to the right hand side of (2.8) for any ergodic measure μ different from the measure concentrated on the sequence of all x (in this case clearly $Z = 2$).

Formula (2.9) follows from

$$S(z_1^n) = n + \sharp\{\alpha \subseteq z_1^n\}$$

and the strong law of large numbers.

Formula (2.10) can be deduced from (2.3).

7.3. $\mathcal{S}\nu$ in terms of ν

We consider the substitution $\alpha \to xy$. We have that

$$W = \lim_{n\to+\infty} \frac{|S(z_1^n)|}{n} = \lim_{n\to+\infty} \sum_{|\underline{z}|=n} \nu(\underline{z})\frac{|S(\underline{z})|}{n},$$

it holds

$$\begin{aligned} \mathcal{S}\nu(\underline{r}) &:= \lim_{n\to+\infty} \tfrac{\sharp\{\underline{r}\subseteq S(z_1^n)\}}{|S(z_1^n)|} = \lim_{n\to+\infty} \tfrac{\sharp\{\underline{r}\subseteq S(z_1^n)\}}{Wn} \\ &= \lim_{n\to+\infty} \tfrac{1}{Wn} \textstyle\sum_{|\underline{z}|=n} \nu(\underline{z})\sharp\{\underline{r}\subseteq S(z_1^n)\}. \end{aligned} \tag{7.25}$$

Suppose now that $|\underline{r}| = k$ and consider, for $n \geq k$,

$$\begin{aligned} D_n &= \textstyle\sum_{|\underline{z}|=n} \nu(\underline{z})\sharp\{\underline{r}\subseteq S(\underline{z})\} - \sum_{|\underline{z}|=n-1} \nu(\underline{z})\sharp\{\underline{r}\subseteq S(\underline{z})\} \\ &= \textstyle\sum_{|\underline{z}|=n} \nu(\underline{z})\mathbb{1}\left(\underline{r} = S(\underline{z})_1^k\right) + \sum_{|\underline{z}|=n-1} \nu(\alpha\underline{z})\mathbb{1}\left(\underline{r} = yS(\underline{z})_1^{k-1}\right). \end{aligned} \tag{7.26}$$

We can rewrite these terms as

$$\begin{aligned} &\textstyle\sum_{|\underline{z}|=n} \nu(\underline{z})\mathbb{1}\left(r = S(\underline{z})_1^k\right) = \sum_{\underline{s}:S(\underline{s})=\underline{r}} \Big(\nu(\underline{s}) + \nu(s_1^{|\underline{s}|-1}\alpha)\mathbb{1}\left(r_k = x\right)\Big) \\ &\textstyle\sum_{|\underline{z}|=n-1} \nu(\alpha\underline{z})\mathbb{1}\left(r = yS(\underline{z})_1^{k-1}\right) = \\ &\qquad\qquad = \textstyle\sum_{\underline{s}:S(\underline{s})=\underline{r}} \Big(\nu(\alpha s_2^{|\underline{s}|}) + \nu(\alpha s_2^{|\underline{s}|-1}\alpha)\mathbb{1}(r_k = x)\Big)\,\mathbb{1}(r_1 = y). \end{aligned} \tag{7.27}$$

Hence D_n is constant for $n \geq k$ and

$$\lim_{n\to+\infty} \frac{1}{W} \sum_{|\underline{z}|=n} \nu(\underline{z})\sharp\{\underline{r}\subseteq S(z_1^n)\} = \frac{1}{W} D_k \tag{7.28}$$

Collecting (7.25)–(7.28) we obtain the expression for $\mathcal{S}\nu$:

$$\begin{aligned} \mathcal{S}\nu(\underline{r}) = \tfrac{1}{W} \quad &\textstyle\sum_{\underline{s}:\,S(\underline{s})=\underline{r}} \Big(\nu(\underline{s}) + \nu(s_1^{|\underline{s}|-1}\alpha)\mathbb{1}\left(r_k = x\right) + \\ &\nu(\alpha s_2^{|\underline{s}|})\mathbb{1}(r_1 = y) + \nu(\alpha s_2^{|\underline{s}|-1}\alpha)\mathbb{1}(r_1 = y)\mathbb{1}(r_k = x)\Big) \end{aligned} \tag{7.29}$$

7.4. $\mathcal{G}\mu$ in terms of μ

The map S inverts G, then in order to find the expression of $\mathcal{G}\mu$ we can invert the expression of $\mathcal{S}\mathcal{G}\mu = \mu$. Let ν be $\mathcal{G}\mu$. The sum on $\underline{s}$ in Eq. (7.29) reduces to $\underline{s} = G(\underline{r})$, namely $\nu(\underline{s}) = 0$ if $\underline{s} \notin G(A^*)$. This reduction make explicitly invertible Eq. (7.29), but we have to distinguish the two cases $x \neq y$ and $x = y$.

Case $x \neq y$.

Let $\underline{r} \in A^*$ and let z, $w \in A$ be such that $z \neq x$ and $w \neq y$. From (7.29) we obtain:

$$\begin{aligned} &W\mu(w\underline{r}z) = \nu(G(w\underline{r}z)) \\ &W\mu(w\underline{r}x) = \nu(G(w\underline{r})x) + \nu(G(w\underline{r})\alpha) \\ &W\mu(y\underline{r}z) = \nu(yG(\underline{r}z)) + \nu(\alpha G(\underline{r}z)) \\ &W\mu(y\underline{r}x) = \nu(yG(\underline{r})x) + \nu(\alpha G(\underline{r})x) + \nu(yG(\underline{r})\alpha) + \nu(\alpha G(\underline{r})\alpha) \end{aligned} \tag{7.30}$$

Let now $\underline{s} = G(\underline{r})$ with $|\underline{s}| = n$ and $|\underline{r}| = k$. The expression of $\nu(\underline{s}) = \mathcal{G}\mu(\underline{s})$ can be calculated from the previous equations, obtaining

$$\begin{aligned} &s_1 \neq y, \quad s_n \neq x: \quad \nu(\underline{s}) = W\mu(\underline{r}) \\ &s_1 = y, \quad s_n \neq x: \quad \nu(\underline{s}) = W(\mu(\underline{r}) - \mu(xyr_2^k)) \\ &s_1 \neq y, \quad s_n = x: \quad \nu(\underline{s}) = W(\mu(\underline{r}) - \mu(r_1^{k-1}xy) \\ &s_1 = y, \quad s_n = x: \quad \nu(\underline{s}) = W(\mu(\underline{r}) + \mu(xyr_2^{k-1}xy) - \mu(xyr_2^k) - \mu(r_1^{k-1}xy)) \end{aligned} \tag{7.31}$$

Now we can calculate $Z = W$ (see Eq. (2.10)) in terms of μ:

$$Z = 1 + \nu(\alpha) = 1 + Z\mu(xy) = \frac{1}{1-\mu(xy)}.$$

We remark that eq.s (7.31) can be synthesized in

$$\mathcal{G}\mu(\underline{s}) = Z \sum_{a,b\in A:\, a\underline{s}b\in G(A^*)} \mu(a\underline{s}b). \tag{7.32}$$

Case $x = y$.

Proceeding as above we obtain again the explicit expressions for $\nu(\underline{s})$ but they are more complicated. As before let $\underline{s} \in G(A^*)$, $|\underline{s}| = n > 0$, $G(\underline{r}) = \underline{s}$, $|\underline{r}| = k$. Let $s_1, s_n \neq x$. Denoting with $\underline{a}^p$ the string of p times the symbol a, the strings in $G(A^*)$ are of the type

$$\underline{\alpha}^p\,\underline{x}^\pi\,\underline{s}\,\underline{\alpha}^q\,\underline{x}^\sigma \quad \text{and} \quad \underline{\alpha}^p\,\underline{x}^\pi, \quad \text{with } p,q \geq 0 \text{ and } \pi,\sigma = 0,\, 1.$$

The expression of $\mathcal{G}\mu = \nu$ in terms of μ is given by:

$$\begin{array}{lll}
\nu(\underline{s}\,\underline{\alpha}^q) & = Z\mu(\underline{r}\,\underline{x}^{2q}) & \text{for } q \geq 0 \\
\nu(\underline{s}\,\underline{\alpha}^q x) & = Z(\mu(\underline{r}\,\underline{x}^{2q+1}) - \mu(\underline{r}\,\underline{x}^{2q+2}))) & \text{for } q \geq 0 \\
\nu(\underline{\alpha}^p) & = Z\sum_{j=0}^{+\infty}(-1)^j\mu(\underline{x}^{2p+j})) & \text{for } p > 1 \\
\nu(\underline{\alpha}^p x) & = Z(\mu(\underline{x}^{2p+1}) - 2\sum_{j=2}^{+\infty}(-1)^j\mu(\underline{x}^{2p+j})) & \text{for } p > 1 \\
\nu(\underline{\alpha}^p\underline{x}^\pi\underline{s}\,\underline{\alpha}^q) & = Z\sum_{j=0}^{+\infty}(-1)^j\mu(\underline{x}^{2p+\pi+j}\underline{r}\,\underline{x}^{2q}) & \text{for } p+\pi \geq 1,\ q \geq 0 \\
\nu(\underline{\alpha}^p\underline{x}^\pi\underline{s}\,\underline{\alpha}^q x) & = Z\sum_{j=0}^{+\infty}(-1)^j\cdot & \\
 & \quad (\mu(\underline{x}^{2p+\pi+j}\underline{r}\,\underline{x}^{2q+1}) - \mu(\underline{x}^{2p+\pi+j}\underline{r}\,\underline{x}^{2q+2})) & \text{for } p+\pi \geq 1,\ q \geq 0
\end{array} \tag{7.33}$$

Now we can calculate Z in terms of μ:

$$Z = 1 + \nu(\alpha) = 1 + Z\sum_{j=0}^{+\infty}(-1)^j\mu(\underline{x}^{2+j}) = \frac{1}{1 - \sum_{j=2}^{+\infty}(-1)^j\mu(\underline{x}^j)}$$

7.5. Proof of Proposition 2.2

This proposition is a consequence of Eq. (7.29) in subsection 7.3, namely $|\underline{s}| \leq \underline{r}$ if $S(\underline{s}) = \underline{r}$.

7.6. Proof of Proposition 2.3

This proposition is a consequence of the fact that the explicit expression (7.29) of $\mu = \mathcal{S}\nu$ in terms of ν can be inverted (in an unique way) if ν respects the pair constraints given by G, as follows from eq.s (7.30)–(7.33) in subsection 7.4. The expression of ν in terms of μ is exactly $\mathcal{G}\mu$, then $\nu = \mathcal{G}\mu = \mathcal{G}\mathcal{S}\nu$.

8. Technical results on entropies transformations

8.1. Proof of theorem 3.1

The result follows from the fact that G is a faithful code and S is a faithful code when restricted to the support of $\mathcal{G}\mu$. We call $C := \{C_n\}_{n\in\mathbb{N}}$ a sequence of universal codes

in the alphabet A and $C' := \{C'_n\}_{n\in\mathbb{N}}$ a sequence of universal codes in the alphabet A' (see theorems II.1.1 and II.1.2 page 122 of [5]).

We have that $C' \circ G$ is a sequence of faithful codes in A. From this we deduce that on a set of μ measure one

$$h(\mathcal{G}\mu) = \lim_{n\to\infty} \frac{C'_{|G(w_1^n)|}(G(w_1^n))}{|G(w_1^n)|} = \lim_{n\to\infty} \frac{n}{|G(w_1^n)|} \frac{C'_{|G(w_1^n)|} \circ G(w_1^n)}{n} \geq Zh(\mu)$$

Likewise we have that $C \circ S$ is a sequence of faithful codes in A'. From this we deduce that on a set of μ measure one

$$h(\mu) = \lim_{n\to\infty} \frac{C_n(w_1^n)}{n} = \lim_{n\to\infty} \frac{|G(w_1^n)|}{n} \frac{C_n \circ S(G(w_1^n))}{|G(w_1^n)|} \geq \frac{h(\mathcal{G}\mu)}{Z}$$

8.2. Proof of theorems 3.2 and 3.3

We proceed splitting the action of G (and then of $\mathcal{G}$) in three parts, introducing two new character b_1, $b_2 \notin A$.
Given a string, we operate as follows:

Step 1 We substitute, starting form the left, any occurrence of xy with xb_1. This operation define a map $R : A^* \to A_R^*$, where $A_R = A \cup \{b_1\}$. We call $\mathcal{R}$ the corresponding map for the measures, defined in the same spirit of Th. 2.2.

Step 2 We substitute any occurrence of xb_1 with b_2b_1. This operation define a map $L : A_R^* :\to A_L^*$, where $A_L = A_R \cup \{b_2\}$. We call $\mathcal{L}$ the corresponding map for the measures.

Step 3 We substitute any occurrence of b_2b_1 with α. This operation, in general, define a map $C : A_L^* :\to A_C^*$, where $A_C = A_L \cup \{\alpha\}$. We call $\mathcal{C}$ the corresponding map for the measures.

From these definitions:

$$C(L(R(\underline{w}))) = G(\underline{w}), \quad \text{and then } \ \mathcal{CLR}\mu = \mathcal{G}\mu.$$

With this splitting we separate the effects of the shortening of the strings (step 3) from the effect of the partial substitutions of characters (steps 1, 2).

Lemma 8.1

$$h_1(\mathcal{R}\mu) \leq h_1(\mu) \tag{8.34}$$

(the proof is in subsection 8.3).

The same assertion holds for $\mathcal{LR}\mu$. Namely we can define L also considering the substitutions starting from the right, namely $x \neq b_1$. In this way $L(\underline{w}) = (R'(\underline{w}^r))^r$, where $\underline{w}^r = (w_1 \ldots w_k)^r = w_k \ldots w_1$ and R' is the substitution, from the left, of b_1x with

$b_1 b_2$. The map R' acts in the same way of R, then lemma 8.1 holds for the corresponding map for the measures $\mathcal{R}'$, and then also for $\mathcal{L}$. In this way we prove that

$$h_1(\mathcal{L}\mathcal{R}\mu) \leq h_1(\mu).$$

The third step preserves h_1 up to the normalization, as stated in the following lemma (proved in subsection 8.4)

Lemma 8.2 *If $\rho \in \mathcal{E}(A_L)$ verifies*

$$\rho(b_2 w) = \rho(z b_1) = 0 \quad \textit{for} \quad w \neq b_1,\, z \neq b_2, \tag{8.35}$$

then

$$h_1(\mathcal{C}\rho) = W h_1(\rho), \tag{8.36}$$

where

$$W = \frac{1}{1 - \rho(b_2 b_1)} = 1 + \mathcal{C}\rho(\alpha). \tag{8.37}$$

We achieve the proof of theorem 3.13 observing that the measure $\rho = \mathcal{L}\mathcal{R}\mu$ verifies the constraints (8.35), then $h_1(\mathcal{G}\mu) \leq W h_1(\mu)$, where $W = Z$ because $W = 1 + \mathcal{C}\rho(\alpha) = 1 + \mathcal{G}\mu(\alpha) = W_\alpha^{\mathcal{G}\mu} = Z_{xy}^\mu$ (see Eq. (2.10)).

We conclude this section remarking that Lemma 8.1 holds also for h_k, and that for h_k it holds the following analogous of Lemma 8.2, proved in subsection 8.5

Lemma 8.3 *Under the hypotheses of lemma 8.2*

$$h_k(\mathcal{C}\rho) \leq W h_k(\rho)$$

From these facts it follows Theorem 3.3.

8.3. Proof of Lemma 8.1

Let $\xi = \mathcal{R}\mu$. The measure μ can be expressed in terms of ξ as follows:

$$\mu(\underline{w}) = \sum_{\underline{z} \colon R(\underline{z}) = \underline{w}} \xi(\underline{z}).$$

We use this formula to express the probabilities of the symbols and of the pairs of symbols.

Case $x \neq y$. Let p be in A:

$$\begin{array}{lll}
\mu(y) = \xi(y) + \xi(b_1), & \mu(p) = \xi(p) & \text{for } p \neq y, \\
\mu(yp) = \xi(yp) + \xi(b_1 p), & \mu(pq) = \xi(pq) & \text{for } p \neq x,\, p \neq y, \\
\mu(xy) = \xi(x b_1), & \mu(xp) = \xi(xp) & \text{for } p \neq y.
\end{array}$$

By direct calculation:

$$\begin{aligned} h_1(\mu) - h_1(\xi) = \; & -\textstyle\sum_{p \in A} (\xi(yp) + \xi(b_1 p)) \log \frac{\xi(yp) + \xi(b_1 p)}{\xi(y) + \xi(b_1)} \\ & + \textstyle\sum_{p \in A} \left(\xi(yp) \log \frac{\xi(yp)}{\xi(y)} + \xi(b_1 p) \log \frac{\xi(b_1 p)}{\xi(b_1)} \right) \end{aligned} \tag{8.38}$$

We prove the lemma showing that:

$$\left(\xi(yp)\log\tfrac{\xi(yp)}{\xi(y)}+\xi(b_1p)\log\tfrac{\xi(b_1p)}{\xi(b_1)}\right)\geq(\xi(yp)+\xi(b_1p))\log\tfrac{\xi(yp)+\xi(b_1p)}{\xi(y)+\xi(b_1)}.$$

Dividing for $\xi(yp)+\xi(b_1p)$ and setting $\beta=\frac{\xi(y)}{\xi(y)+\xi(b_1)}$, $\gamma=\frac{\xi(yp)}{\xi(yp)+\xi(b_1p)}$, this inequality can be rewritten as:

$$\gamma\log\frac{\beta}{\gamma}+(1-\gamma)\log\frac{1-\beta}{1-\gamma}\leq 0,$$

which is always verified.

Case $x=y$. Let $p\in A$, $p\neq x$:

$$\begin{array}{ll}\mu(x)=\xi(x)+\xi(b_1), & \mu(p)=\xi(p),\\ \mu(xx)=\xi(xb_1)+\xi(b_1x), & \mu(xp)=\xi(xp)+\xi(b_1p),\\ \mu(pq)=\xi(pq)\quad\text{for } q\in A, & \mu(px)=\xi(px).\end{array}$$

The difference between the 1–conditional entropies is

$$\begin{aligned}h_1(\mu)-h_1(\xi)= \; & -\textstyle\sum_{p\in A,\,p\neq x}(\xi(xp)+\xi(b_1p))\log\frac{\xi(xp)+\xi(b_1p)}{\xi(x)+\xi(b_1)}\\ & -(\xi(xb_1)+\xi(b_1x))\log\tfrac{\xi(xb_1)+\xi(b_1x)}{\xi(x)+\xi(b_1)}\\ & +\textstyle\sum_{p\in A,\,p\neq x}\left(\xi(xp)\log\frac{\xi(xp)}{\xi(x)}+\xi(b_1p)\log\frac{\xi(b_1p)}{\xi(b_1)}\right)\\ & +\xi(xb_1)\log\tfrac{\xi(xb_1)}{\xi(x)}+\xi(b_1x)\log\tfrac{\xi(b_1x)}{\xi(b_1)}\end{aligned}\tag{8.39}$$

We prove that this difference is positive with the same argument used for the case $x\neq y$.

Finally, we remark that in the same way we can prove that $h_k(\xi)\leq h_k(\mu)$.

8.4. Proof of Lemma 8.2

Let $\nu=\mathcal{C}\rho$ and $W=1+\nu(\alpha)$. It is easy to write ρ in terms of ν. Let $p,q\neq b_1,b_2$. The probabilities of the symbols and of the pairs of symbols are given by

$$\begin{array}{ll}W\rho(b_1)=W\rho(b_2)=\nu(\alpha) & W\rho(p)=\nu(p)\\ W\rho(pb_1)=W\rho(b_2q)=0 & W\rho(pq)=\nu(pq)\\ W\rho(pb_2)=\nu(p\alpha) & W\rho(b_1q)=\nu(\alpha q)\\ W\rho(b_2b_1)=\nu(\alpha) & W\rho(b_1b_2)=\nu(\alpha\alpha)\end{array}$$

By explicit calculation:

$$\begin{aligned}H_1(\rho) &= -\textstyle\sum_{p\in A_C\setminus\alpha}\frac{\nu(p)}{W}\log\frac{\nu(p)}{W}-2\frac{\nu(\alpha)}{W}\log\frac{\nu(\alpha)}{W}=\frac{H_1(\nu)}{W}+\frac{\log W}{W}-\frac{\nu(\alpha)}{W}\log\frac{\nu(\alpha)}{W},\\ H_2(\rho) &= -\textstyle\sum_{p,q\in A_C}\frac{\nu(pq)}{W}\log\frac{\nu(pq)}{W}-\frac{\nu(\alpha)}{W}\log\frac{\nu(\alpha)}{W}=\frac{H_2(\nu)}{W}+\frac{\log W}{W}-\frac{\nu(\alpha)}{W}\log\frac{\nu(\alpha)}{W}.\end{aligned}$$

Then

$$h_1(\rho)=H_2(\rho)-H_1(\rho)=\frac{h_1(\nu)}{W}.$$

8.5. Proof of Lemma 8.3

We need some definitions. Let $\underline{w} = w_1^l$. We can identify $\underline{w}$ with the cylindrical subset $K_{\underline{w}} \subseteq A^{\mathbb{Z}}$ defined as follows

$$K_{\underline{w}} = \{\underline{x} \in A^{\mathbb{Z}} : x_{-l} = w_1, x_{-l+1} = w_2, \ldots, x_{-1} = w_l\}$$

Let $P \subseteq A^*$ be a finite set. We say that P is a partition if

$$\{K_{\underline{w}}\}_{\underline{w}\in P} \quad \text{is a partition of } A^{\mathbb{Z}}, \text{ i.e. } \begin{cases} (1) & K_{\underline{w}} \cap K_{\underline{z}} = \emptyset \ \text{ if } \underline{w} \neq \underline{z}, \\ (2) & \bigcup_{\underline{w}\in P} K_{\underline{w}} = A^{\mathbb{Z}}. \end{cases}$$

Condition (1) says that any string of P is not a suffix for others strings of P. If only condition (1) is verified, we say that P is a semi-partition. It is easy to show that any semi-partition can be completed to obtain a partition. Moreover, if the minimum of the length of the strings in P is l, we can complete P using strings of length greater or equal to l.

If P is a partition, we can define the P-conditional entropy as

$$h_P(\mu) = -\sum_{\underline{w}\in P,\, a\in A} \mu(\underline{w}a) \log \frac{\mu(\underline{w}a)}{\mu(\underline{w})}.$$

If P and Q are two partition we say that P is more fine of Q if any string of P ends with a string of Q. If P is more fine then Q:

$$h_P(\mu) \leq h_Q(\mu). \tag{8.40}$$

(The proof is at the end of this section).

Note that

$$P = \{\underline{s} \in A_L^* |\, |C(\underline{s})| = k\},$$

is a semi-partition, and that, from direct calculation

$$h_k(\nu) = W h_P(\rho).$$

Where we remember that $\nu = \mathcal{C}\rho$. In particular we have used that, if $\underline{s} \in A_L^*$, $\rho(\underline{s}b_2) = \rho(\underline{s}b_2b_1)$ and if the last symbol of $\underline{s}$ differs from b_2 then $\rho(\underline{s}b_1) = 0$.

Finally let $\overline{P}$ be a completion of P.

$$h_k(\nu) = W h_P(\rho) \leq W h_{\overline{P}}(\rho).$$

The length of the strings in P is greater or equal to k and we construct $\overline{P}$ so that the same holds for $\overline{P}$. Therefore, A_L^k is a partition less fine of $\overline{P}$. Invoking Eq. (8.40) we conclude that

$$h_k(\nu) \leq W h_{A_L^k}(\rho) = W h_k(\rho).$$

Proof of Eq. (8.40).

Let $\underline{w} \in Q$ and $X_{\underline{w}} \subseteq P$ be the subset of the strings which end with $\underline{w}$. From this definition:

$$P = \bigcup_{\underline{w} \in Q} X_{\underline{w}}, \quad \mu(\underline{w}) = \sum_{\underline{r} \in X_{\underline{w}}} \mu(\underline{r}).$$

The function $\Phi(x) = x \log x$ is convex, then if $\lambda_i \geq 0$ and $\sum \lambda_i = 1$, $\Phi(\sum \lambda_i x_i) \leq \sum \lambda_i x_i \log x_i$. Now

$$-h_Q(\mu) = \textstyle\sum_{\underline{w} \in Q} \mu(\underline{w}) \sum_{a \in A} \mu(a|\underline{w}) \log \mu(a|\underline{w}),$$

and

$$\mu(a|\underline{w}) = \tfrac{\mu(\underline{w}a)}{\mu(\underline{w})} = \textstyle\sum_{\underline{r} \in X_{\underline{w}}} \frac{\mu(\underline{r}a)}{\mu(\underline{w})} = \sum_{\underline{r} \in X_{\underline{w}}} \frac{\mu(\underline{r}a)}{\mu(\underline{r})} \frac{\mu(\underline{r})}{\mu(\underline{w})}.$$

Indicating with $x^a_{\underline{r}} = \mu(\underline{r}a)/\mu(\underline{r})$, and with $\lambda_r = \mu(\underline{r})/\mu(\underline{w})$ and noting that $\sum_{\underline{r} \in X_{\underline{w}}} \lambda_{\underline{r}} = 1$, we obtain:

$$\begin{aligned} -h_Q(\mu) &= \textstyle\sum_{\underline{w} \in Q} \sum_{a \in A} \mu(\underline{w}) \Phi\left(\sum_{\underline{r} \in X_{\underline{w}}} \lambda_{\underline{r}} x^a_{\underline{r}}\right) \\ &\leq \textstyle\sum_{\underline{w} \in Q} \sum_{\underline{r} \in X_{\underline{w}}} \sum_{a \in A} \mu(\underline{w}) \frac{\mu(\underline{r})}{\mu(\underline{w})} \mu(a|\,\underline{r}) \log \mu(a|\,\underline{r}) \\ &= \textstyle\sum_{\underline{r} \in P} \sum_{a \in A} \mu(\underline{r}) \mu(a|\,\underline{r}) \log \mu(a|\,\underline{r}) = -h_P(\mu) \end{aligned} \tag{8.41}$$

Acknowledgements

We thank V. Loreto and G. Parisi for very useful discussions. D.B. acknowledges the support of cofin MIUR. E.C. acknowledges the support of GNFM-INdAM and cofin MIUR. D.G. acknowledges the support of GNFM-INdAM and cofin prin prot. 2004028108_005.

References

[1] J. Adamek *Foundations of coding* John Wiley and Sons (1991).

[2] Argenti F., Benci V., Cerrai P., Cordelli A., Galatolo S., Menconi G. *Information and dynamical systems: a concrete measurement on sporadic dynamics. Classical and quantum complexity and non-extensive thermodynamics* (Denton, TX, 2000), Chaos Solitons Fractals **13**, (2002), no 3, 461-469.

[3] W. Ebeling, M.A. Jiménez–Montaño *On grammars, complexity, and information measures of biological macromolecules* Math. Biosc **52** 53–71 (1980)
M.A. Jiménez–Montaño *On the syntactic structure of protein sequences and the concept of grammar complexity* Bull. Math. Bio. **42** 641–659 (1984)
M.A. Jiménez–Montaño, W. Ebeling, T. Pöschel *SYNTAX: A computer program to compress a sequence and to estimate its information content* ArXiv:cond-mat/0204134
P.E. Rapp, I.D. Zimmermann, E.P. Vining, N. Cohen, A.M. Alabano, M.A. Jiménez–Montaño *The algorithmic complexity of neural spike trains increases during focal seizures* The Jour. of Neuroscience **14** (8) 4731–4739, and Phys. Lett. A **192** 27–33 (1994)

[4] P. Grassberger *Data compression and entropy estimates by non-sequential recursive pair substitution* ArXiv:physics/0207023

[5] P.C. Shields *The ergodic theory of discrete sample paths* Graduate Studies in Mathematics **13**, American Mathematical Society (1996).

[6] T. A. Welch *A technique for high performance data-compression* IEEE Computer, 17(6), 8-19, 1984.

[7] J. Ziv, A. Lempel *A universal algorithm for sequential data compression* - IEEE Transactions on Information Theory, 23(3), 337-343, 1977.

[8] J. Ziv, A. Lempel *Compression of individual sequences via variable-rate coding* - IEEE Transactions on Information Theory, 24(5), 530-536, 1978.